Assessing language learning in virtual exchange: suggestions from the field of language assessment

Jiyoon Lee[1] and Shannon Sauro[2]

Abstract

The goal of this paper is to enhance Virtual Exchange (VE) practitioners' language assessment literacy. To do so, it begins with an overview of assessment practices commonly used in VE for evaluating the complex and multifaceted nature of language competence. These include the following: (1) approaches that evaluate change in learners' language use over time, (2) approaches that employ pre- and post-tests to evaluate learning outcomes, and (3) approaches that rely on students' self-report or self-documentation of learning. Based on this overview, we then look to the field of language assessment for guidelines on the selection and use of classroom-based and standardized assessment tools and practices. Using an existing VE practice scenario, the authors provide an example of language assessment selection and development.

1. Introduction

One of the practical challenges facing the implementation of a complex teaching practice such as VE is providing sufficient, measurable, rigorous, and compelling evidence of learning for stakeholders

1. University of Maryland, Baltimore County, USA; jiyoon@umbc.edu

2. University of Maryland, Baltimore County, USA; ssauro@umbc.edu

including researchers, practitioners, and learners. University stakeholders and funding agencies necessarily require such evidence when deciding whether to invest money and resources in the training and implementation of VE; researchers who seek to evaluate the influence of certain VE practices on student beliefs, knowledge, or behaviors require specific tools to evaluate these changes, and practitioners, particularly those who are new to VE, are often in need of tools to assess students' learning of skills and knowledge in alignment with course and program goals. Further complexifying the practice of assessing VE is identifying tools and measures appropriate for the wide range of skills, knowledge, and behaviors that VE supports. For instance, large-scale funded European projects have been developed with the intent of evaluating learning outcomes related to intercultural communicative competence, digital and critical digital literacies, disciplinary skills, and language competence (e.g. EVOLVE Project Team, 2020; The EVALUATE Group, 2019). In addition to these skill areas, outcomes from the European Commission's pilot project, Erasmus+ VE, (Helm & van der Velden, 2019) and research guidelines from the US-based Stevens Initiative (The Stevens Initiative, 2019) highlight global competences or other 21st century skills such as cross-cultural communication and collaboration, empathy, tolerance, critical thinking, and problem-solving.

Taken together, VE researchers and practitioners face a monumental task in identifying, purchasing, or even developing the methods and instruments needed to adequately assess relevant student learning. Accordingly, this paper sets out to provide an overview of research on assessment practices commonly used in VE for evaluating the complex and multifaceted area of language competence and looks to the field of language assessment for guidelines on the selection and use of informal and standardized assessment tools and practices. By doing so, another aim of the present paper is to enhance VE practitioners' language assessment literacy. Language assessment literacy refers to stakeholders' understanding of language assessment principles and abilities to select, identify, and design appropriate language assessment, and use language assessment and its results (Fulcher, 2012; Inbar-Lourie, 2008; Lee, 2019; Lee & Butler, 2020).

2.   Literature review: language assessment in VE

VE, historically known as telecollaboration in the field of computer-assisted language learning, has a long history as a pedagogical practice to support foreign language development (O'Dowd, 2016). Research on the use of telecollaboration/VE for foreign language learning has a rich history, dating back approximately 25 years (e.g. Cummins & Sayers, 1995; Little & Brammerts, 1996). Accordingly, language learning is one of the richest skill areas within VE to be the object of varied assessment practices. This can be seen in Lewis and O'Dowd's (2016) systematic review of research

studies which examined foreign language learning through VE, and which served as a starting point for this paper and our categorization of different types of language assessment practices used in the VE literature.

Lewis and O'Dowd's (2016) systematic review of studies, which examined learning, including foreign language learning through VE carried out at the university level, provided an overview of commonly used assessment approaches in studies of individual VEs. Of the 54 studies reviewed, the authors identified 24 that specifically focused on the development of foreign language skills. As these 24 studies were smaller in scale, usually involving learner populations from two partner classes and thus reflecting assessment practices feasible to smaller groups, we also reviewed the results of more recent large-scale multi-site projects on VEs which also set out to assess language development (i.e. EVALUATE, EVOLVE). In addition, we looked to more recent literature published on telecollaboration or VE which foregrounded questions of second language acquisition development to identify studies that also used standardized language assessments. The following sections describe the three main approaches to assessing language learning we identified through this process and provide an overview of key studies which illustrate the different types of assessment practices encompassed by each approach.

2.1. Approaches that evaluate changes in language use over time during the VE

The first set of approaches for assessing language learning has in common an emphasis on collecting and examining learner discourse generated at different times during the VE. This encompasses studies which Lewis and O'Dowd (2016) describe as using transcript analysis because the tight timeframe involved in this type of assessment usually draws upon analysis of discourse that is written and may be gathered from discussion board postings, chatlogs, or emails. This set of approaches include the following: (1) corpus-based analysis of learners' target language use over time (e.g. Belz & Vyatkina, 2005, 2008), and (2) multiple measures combining learner self-reports and subsequent use of forms (recycling) during the VE (e.g. Vinagre & Muñoz, 2011).

2.1.1. Corpus-based analysis of learners' target language use over time

Corpus-based approaches to assessing learner language rely on the collection and analysis of corpora, which are large collections of machine-readable authentic texts. In this case the authentic texts collected are the students' own written interactions generated during the VE (e.g. online discussion forums, emails, and even text chats). Through the use of concordancing programs, defined as "text search engines with sorting functions" (Cotos, 2017, p. 249; see also Anthony, 2019, for an example of a freely available concordancer known as AntConc), instructors and researchers can categorize

and analyze the corpus of learner generated language to identify trends and patterns in language development and mastery of specific target language forms.

This approach is exemplified in several studies which investigated L2 German learners' use of several pragmatic features (German modal particles *ja*, *denn*, *doch*, *mal*, and the German *da*-compound) during a trans-Atlantic partnership pairing university students in the United States and Germany (Belz & Vyatkina, 2005, 2008). The corpus consisted of students' emails and chats which were compiled while the exchange was on-going and used as both instructional intervention (i.e. L2 learners had the opportunity to analyze their use of these particular features and compare it with their German partners' use) and as an assessment of their use of the form. Specifically, the researchers tabulated the frequency of the targeted forms used by the learners (both total occurrence and as a ratio of total words produced) pre- and post-teaching intervention, and compared these results with the relative frequency of the target forms used by their German-speaking peers. Findings revealed a marked increase in target form use from pre- to post-intervention, including a mixture of both accurate and inaccurate application (overuse). The advantage of this type of assessment of learner language is that it provides a longitudinal look at the development of language features over time. Limiting the practicality of this approach to assessing language learning, however, is the increasing use of synchronous video-conferencing tools incorporated into VEs, necessitating an extremely time intensive transcription of recordings in order to compile a corpus. In addition, recordings of learner interactions are not always permitted in VEs due to data protection concerns in various educational contexts or regions.

2.1.2. Multiple measures combining learner self-reports and subsequent use during the VE

A second approach that evaluates language development change over time during a VE is exemplified by Vinagre and Muñoz (2011) who reported on a three-month long tandem exchange in which participants used email to alternate the target languages of the two partner classes: L2 Spanish for the students in Germany, and L2 German for the students in Spain. During this partnership, emails were written half in German and half in Spanish, eliciting peer feedback on grammar errors. To evaluate learning, this approach incorporated multiple measures including error recycling in subsequent emails (i.e. L2 learners' incorporation of the vocabulary and grammar their L1 partners had corrected in prior emails) as well as learner self-reports of learning in a language learner diary that students kept during the VE. These two primary measures were augmented by self-evaluation measures and individual interviews. The advantage of using multiple measures, including self-evaluation, was that it allowed researchers and teachers the opportunity to focus on focal moments in the email interactions to identify the specific items each learner received feedback on, thereby allowing for an individualized evaluation of learner language development over time. However, such

individualized attention to each learner's performance is also labor intensive and may be beyond the scope of what most language teachers overseeing a VE while also teaching non-VE elements of the course have time for.

2.2. Approaches that employ pre- and post-tests to evaluate learning outcomes

The second set of approaches are those which most resemble prototypical forms of assessment in that they rely on the use of tests given before and after the VE to evaluate changes in language knowledge or language proficiency. These types of assessments are typically carried out by researchers interested in evidence that the VE itself or some particular teaching intervention during the VE is responsible for language development (as opposed to other language learning experience L2 learners may have encountered while the VE was running). These approaches differ in the specific type of pre- and post-test used and include teacher/researcher designed assessment of a specific language feature or a standardized language assessment of broader language skills.

An example of the former can be seen in the study by Sauro (2009), which examined a VE carried out between university students in the United States and university students in Sweden training to be secondary school English teachers. In pairs, Swedish and US students collaborated in English on two writing activities meant to elicit errors with the zero article with abstract noncount nouns, a linguistic form which has proven challenging for even high proficiency Swedish learners of English to master. While the Swedish students provided the content knowledge for the writing activities (e.g. Swedish culture, environmentalism), their US peers provided language expertise and had been instructed to give specific types of feedback in response to errors with the zero article. Thus, as in Vinagre and Muñoz (2011) discussed above, the primary focus of this study was on learning that arose in response to peer feedback. Learning was assessed using acceptability judgment tests to measure knowledge of the English zero article; in other words, L2 learners were asked to determine if sentences were grammatically acceptable or unacceptable. The study followed a pre-, post-, and delayed post-test design in which each test contained 35 items, of which 15 targeted the zero article with abstract noncount nouns and 20 were distractor items. All tests and items had been piloted and refined during a preliminary study with a similar population. As Sauro's (2009) study illustrates, one of the advantages of using tests specifically designed to measure learning in this manner is the ability to focus on the development of skills and knowledge related to a predetermined language form, which may otherwise not be captured in more naturalistic types of assessment if it is otherwise easy to avoid in communication. On the other hand, this type of assessment is more ideal for research studies than for actual classroom practice since it demands a high degree of time and opportunity to develop and only focuses on a limited range of items or knowledge.

The other type of assessments found in studies that employ pre- and post-test designs addresses these limitations by relying on pre-existing tests that have already been developed to assess language skills and knowledge more broadly, for instance standardized tests or components of standardized tests. An example of this can be seen in Saito and Akiyama's (2018) investigation of the English spoken language development by Japanese learners of English following a semester-long VE with native English-speaking partners that was carried out using video-conferencing. Language learning was assessed with pre- and post-tests that consisted of portions of the listening component of two versions of the Test Of English for International Communication (TOEIC), an international standardized test used to assess English language proficiency. Specifically, these components included the following three types of listening tasks: (1) short question and response sequences to assess basic listening proficiency, (2) conversations to assess listening comprehension in conversational situations with a high degree of turn-taking, and (3) longer sustained listening sequences such as those found in talks to assess listening comprehension of linguistically and semantically complex input.

Much as in Sauro (2009), this study employed a quasi-experimental design using pre- and post-tests along with treatment and comparison groups to evaluate the influence of feedback during VE on language development. However, since Saito and Akiyama (2018) focused on the broader skill of listening comprehension and not on specific language forms that were particular to a certain learner population, they did not need to develop a customized measure of learning and could instead rely on a standardized assessment. Despite this advantage, standardized tests often come with a high price tag and may not be affordable for many practitioners or researchers who may have limited funds. An additional complication can arise in the case of large-scale, multi-site studies which explore language development across various contexts and in various languages, for which no single standardized measure will suffice. This issue is addressed in the final set of approaches to assessment.

2.3.   Approaches that rely upon student self-assessment or self-documentation

The third set of approaches discussed here, student self-assessment or self-documentation of language learning, can be found in both small-scale practitioner-oriented reports that rely on assessment tools that have been designed in accordance with course goals as well as in large-scale multi-site studies whose large and linguistically varied learner populations make other forms of assessment unfeasible or too time-consuming to implement.

An example of this found in a small-scale single partnership study can be seen once again in Vinagre and Muñoz (2011) in the language learning diaries that students kept throughout the VE and in which they documented their language learning experiences. In large-scale multi-site studies, such

learner self-documentation may be more likely to take the form of a portfolio, which requires students to write responses to specific prompts intended to elicit reflection on language learning experiences at different points during the VE. One such example can be found in the data collection of the EVOLVE project (https://evolve-erasmus.eu/research-details/), a large multi-site VE project designed to examine student learning outcomes across a range of disciplines.

However, the use of portfolios and diaries, while resulting in rich student reflections and self-reports of language learning, also generates an overwhelming amount of data in large-scale studies. For this reason, more easily analyzed quantitative self-assessment measures are also used such as self-evaluative can-do statements using Likert scales which students complete both pre- and post-VE. An example of one of nine such items used in the pilot round of the EVOLVE project (EVOLVE Project Team, 2020) is shown in Table 1.

Table 1. Example can-do item type (from EVOLVE Project Team, 2020, p. 84)

Item	Language skill or knowledge
I can use the precise words needed to communicate specific meaning or to express the exact meaning of what I want to say.	Vocabulary control

This specific item is one of several used in this project which draws upon the CEFR[3] Illustrative Descriptor Scales (Council of Europe, 2018) to reflect the skills and knowledge identified as necessary for effective communication in a second or foreign language (Council of Europe, 2001). As already mentioned, one of the benefits of such a simple self-assessment item type for large-scale studies is that it can quickly be evaluated. However, one of the limitations of such an approach, particularly for VE practitioners, lies in the time and skill required to develop and pilot items to ensure effectiveness among specific learner populations (see for example, The EVALUATE Group, 2019, for a detailed description of the process entailed in developing similar items to evaluate intercultural competence and digital skill development through VE for language teacher candidates).

3.   Language assessment guidelines for VE practitioners

As noted earlier, varied language assessments have been implemented in VE practices to hold VE stakeholders accountable and collect evidence of VE practices' effectiveness in developing learners' target language. While assessment-related tasks are prominent in VE practices, VE practitioners often still require valid yet reasonable and affordable measures of language learning which are relevant

---

3. Common European Framework of Reference for languages

to their learner populations and thus require more information about identifying, designing, and using language assessments. In this section forward, we attempted to provide tools for such tasks.

3.1.   Starting point

The starting point of language assessment selection or development for VE practices is to revisit the objectives of the particular VE practices. The objectives of a VE will guide the stakeholders to decide the target constructs, formats and methods, and agents of assessment. Selecting adequate language assessment that provides constructive information about learners to stakeholders begins with identifying and operationalizing target constructs. Constructs are an ability or set of abilities, skills, and traits that are derived from theories and that are not directly observable unless they are intentionally elicited for evaluation (Bachman & Palmer, 2002). Constructs should be driven by the objectives of VE courses, modules, or lessons. In order to determine constructs, stakeholders need to think about questions like *what skills or abilities did I/we teach via the VE practice?*, *what is the goal of this VE practice?*, and *what do I/we need to assess?* Operationalization of constructs can start from identifying proto skills such as grammatical competence, sociolinguistic competence, strategic competence, and global communication. Depending on the learning goals, VE practitioners can decide to assess receptive skills (i.e. listening and reading abilities), productive skills (i.e. speaking and writing abilities), or integrated abilities (i.e. combination of both receptive and productive skills). Then, VE practitioners can narrow down the construct definition with appropriate specification (Carr, 2011). If the construct definition is too broad, the assessment to select or to develop will not have a clear direction. If the construct is too narrowly defined, the assessment will not be comprehensive enough to help stakeholders make informed decisions about learners. Constructs will also be reflected in prompts on language assessment as well as in a rubric. Each criterion found in a rubric is closely related to specific constructs or their subconstructs. We provided examples in the later section.

It is also important for stakeholders to decide the *purposes* of language assessment. VE stakeholders can use language assessment to place learners by identifying their current proficiency levels not related to a particular curriculum or unit (i.e. proficiency assessment), measure their achievement after they learned a curriculum/unit (i.e. achievement assessment), or diagnose learners' strengths and weaknesses to provide adequate remedial support (i.e. diagnostic assessment). The assessment purposes and the constructs should be closely connected.

The agents of language assessment in VE may include educators, learners, administrators, or parents and whole communities. The degree of each agent's involvement in assessment selection and development can also impact the types of language assessment in VE practices. It is also necessary to

map out the physical locations where assessment takes place and the ways to implement assessment including the interface of assessment (i.e. audio, video-conferencing, text-chat), and the number of assessment participants (i.e. whole class, small groups, paired, or self-assessment). VE practitioners also need to set up a clear expectation of the way assessment is presented, learners' performance is evaluated, the assessment results are shared, and relevant standards.

3.2. Language assessment types

Once VE practitioners set up a language assessment blueprint based on the previously mentioned factors, VE practitioners can decide the types of language assessment to select or develop. The ranges of language assessment used in the VE studies reviewed in the previous sections can be largely categorized based on its relation to the VE curriculum (Figure 1): curriculum-based assessment and curriculum-free assessment. Curriculum-based assessment includes formative and summative assessment, and curriculum-free assessment includes proficiency, diagnostic, and placement assessment. The types of language assessment presented in Figure 1 are not exhaustive; however, they help support a general understanding of varied kinds of language assessment.

Figure 1. Examples of language assessment
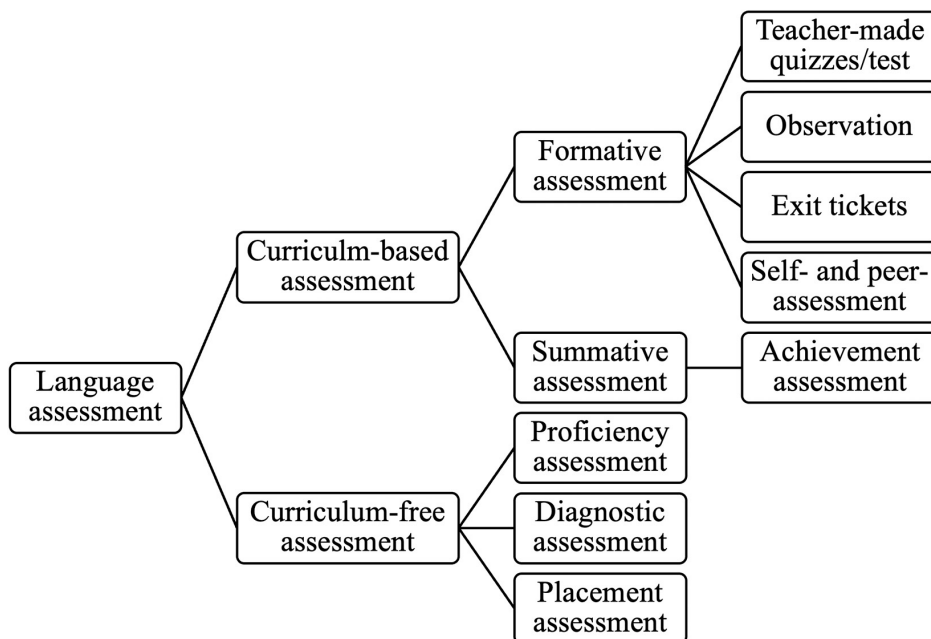
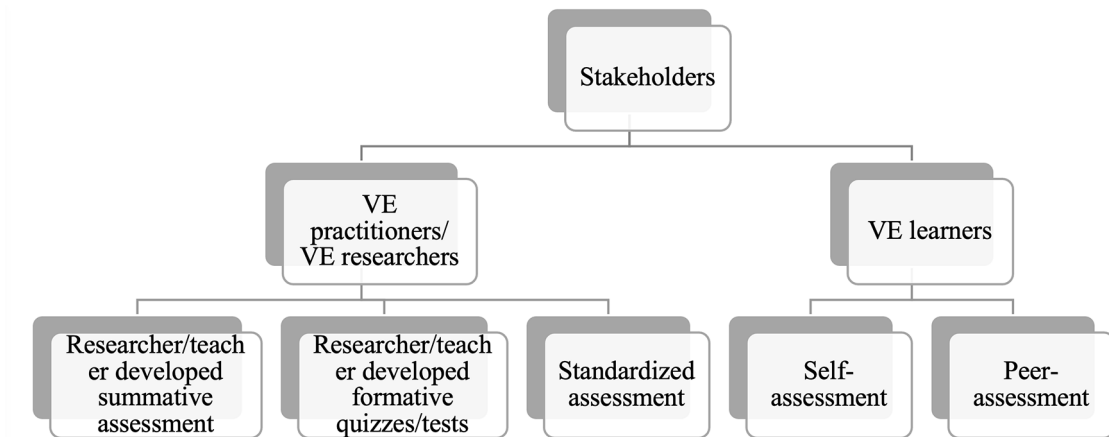### 3.2.1.   Curriculum-based language assessment

Curriculum-based language assessment reflects specific learning goals that are closely related to a particular curriculum/course. Depending on the ways to use assessment results, curriculum-based language assessments can be categorized into formative or summative assessment. Through on-going formative assessment, VE practitioners collect information in order to use it for the next implementation of VE practices. Using assessment results for formative purposes, the results help VE students understand their strengths and weaknesses in L2 learning and help VE instructors prepare the next lesson (Belz & Vyatkina, 2005, 2008). Varied types of formal or informal formative assessment are available (Lee, 2020; Solano-Flores, 2016). Formal formative assessment is planned in advanced by VE practitioners and involves everyone who participates in VE practices. Informal formative assessment is individualized and in-situ assessment practices including using eye-contact to check learners' comprehension and unplanned question-and-answers. If VE practitioners implement language assessment results to confirm the pedagogical effectiveness or conclude the current cycle of the VE practices, this type of assessment can be used for summative purposes. Assessment results can also be used for research purposes only or come at the end of the instructional cycle (i.e. summative assessment). An example of summative assessment includes an achievement assessment at the end of a unit, a lesson, the mid-term, or the final. To differentiate formative from summative assessment, studies that incorporated on-going corpus analysis used a form of formative assessment if the VE practitioners used the results to change the following iteration of VE practices. In contrast, the language assessment implemented as a post-test in Sauro (2009) which measured VE participants' comprehension and uses of the English zero article was curriculum-based, summative assessment.

3.2.2.   Curriculum-free language assessment

Some language assessments were not specifically related to a particular VE curriculum but used to measure VE participants' general language proficiency before or after VE practices. Curriculum-free assessment includes proficiency assessment, placement assessment, and diagnostic assessment. These types of language assessments are not linked to a particular curriculum. For instance, Saito and Akiyama (2018) used the TOEIC to assess the differences in VE students' L2 proficiency before or after the VE. They did not attend to any particular linguistic information provided through the VE practices but instead measured general proficiency development as a result of the VE practices. The pre-test used in Sauro's (2009) study could be considered a diagnostic test if she had implemented it to understand her participants' current knowledge of target features and planned her task materials based on the assessment results.

Language assessments can also be categorized based on the stakeholders who play a leading role in language assessment practices. Figure 2 shows different stakeholders and varied types of language assessments.

Figure 2.   Stakeholders of language assessment

When VE researchers or practitioners lead assessment practices, they can choose to design curriculum-based language assessment which reflects VE practices directly or select a standardized language assessment available on the market. In Appendix 1, we presented examples of standardized language assessments and relevant information including cost, target language and test-takers, and turnaround time. Standardized language assessment is language assessment that is norm-referenced and is designed to make a comparison among test-takers. By definition, standardized language assessment should be implemented and evaluated in a standardized and consistent way to make comparison meaningful. Standardized language assessment would be appropriate when a large-scale VE practice is implemented, and the stakeholders intend to compare learners' performance across groups.

When VE learners play a leading role in VE language assessment, self- and peer-assessment are good options. If it is well-designed and implemented, self-assessment can help learners actively monitor their own learning (Butler & Lee, 2010; Harris, 1997). It is also possible that peer-assessment can promote learners' higher order thinking and critical observation skills (Brown, 2004; Cheng & Warren, 2005). Mok (2011) suggested that peer-assessment can have benefits including learners'

increased sense of responsibilities, enhanced metacognition, abilities to accurately evaluate, and active engagement in learning. Both self- and peer-assessment invite learners to act as autonomous and responsible agents to evaluate their own and their peers' performance in VE practices. Self-assessment is usually constructed using 'can-do' statements, which helps learners assess their capabilities of completing a particular task. The self-report used in the EVOLVE Project Team's (2020) research was an example of self-assessment using a can-do statement. Can-do statements can be derived from a set of standards and course/lesson objectives. Research has shown that teachers can expect more accurate results when concrete wording is used in self-assessment statements, clearly defined behavior or performance is assessed, and self-assessments are implemented right after the task the learners are supposed to assess (Butler & Lee, 2006).

4.   Practical suggestions

When identifying or developing a language assessment for VE practices, we suggest considering the below.

- List the objectives of a course or a lesson if you want to design an assessment that is curriculum-based. If you want to design a language assessment that is free from a particular curriculum, state a goal to accomplish with the VE practice in general.
- Review the objectives and identify the possible constructs that can be elicited and observed for assessment.
- Define constructs to include in the assessment.
- Determine assessment types, agents, interface, an evaluation plan, and an assessment results dissemination plan.
- If learners' performance is part of constructs, VE practitioners need to develop a rubric. It is also possible to invite VE learners to design a rubric together.

To put the guideline into practice, we adopted one of the VE practices implemented in a German class below. In the VE practice, the learners spoke French and Czech as their first language, and their target language (i.e. German) proficiency was A2 or B1 level of the CEFR. The learners interacted with their VE partners in Czechia to learn about their home country. In the first class, the instructor set three lesson objectives as noted in the first column of Table 1. Adopting her lesson objectives, we identified possible constructs for an assessment (second column) and proposed examples of assessments (fourth column). The third column shows the stakeholder of the assessment and the ways to evaluate learners' performances as well as disseminate the assessment results.

Table 2. **Example of objectives-constructs-assessment relationship**

Objectives	Constructs	Example assessment	
get to know each other better (using related questions / answers + adequate pronouns) => exchange basic information (who we are, where we live, study, …)	Students' ability to obtain information about their VE partners by using accurate question forms and pronouns in target language	• Evaluation agent: VE instructor • Interface: Google Docs/Form • Evaluation: pay attention to meaning • Result: return evaluation upon next class	• Fill-in-the blanks with the information they obtained from their VE partners
		• Evaluation agent: VE instructor • Interface: Google Docs/Form • Evaluation: pay attention to forms • Result: return evaluation upon next class	• A paragraph description of VE partners
be able to describe his/her environment (large + narrow)	Students' ability to describe their environment by using contrastive adjectives	• Evaluation agent: VE learners • Interface: Google Form – quiz format • Evaluation: automatic evaluation • Result: share the results with the learners	• Checklist that includes examples of contrastive adjectives
		• Evaluation agent: VE practitioner • Interface: in-class presentation • Evaluation: rubric that reflects constructs • Result: rubric	• Oral report of their VE partners' home country using contrastive adjectives
maintain the conversation => be able to ask for explanations / additional information / indicate that you have understood or not understood), ask for repetition, be able to rephrase	Students' ability to ask clarification questions, explanation, and repetition from their VE partners	• Evaluation agent: VE learners • Interface: Google Form • Evaluation • Result	• Self-assessment that asks how many and what types of questions they asked to their VE partners to maintain the interaction
		• Evaluation agent: VE learners • Interface: VE environment • Evaluation: Rubric • Result: share each other's evaluation using the rubric	• Peer-assessment that asks VE learners to evaluate (1) each other's uses of target language forms, (2) success of using clarification questions, repetition, and explanation request

The first lesson objective is "get to know each other better (using related questions / answers + adequate pronouns) => exchange basic information (who we are, where we live, study, …)". This objective can be divided into (1) getting information about VE partners and (2) using certain

grammatical forms. Hence, the construct can be students' ability to obtain information and use proper grammatical forms. In this example, we define the construct as "students' ability to obtain information about their VE partners by using accurate question forms and pronouns in target language". Examples of assessment to measure the construct included fill-in-the blanks and a paragraph writing.

The example we provide in Table 2 shows assessment for each lesson objective. However, VE practitioners can choose to design one or two assessments to measure the constructs. That is, VE practitioners can use brief oral presentations in which VE students report the information that they found from their VE partners and complete self-assessment about their language use during the interaction with their VE partners. Using self-assessment, VE practitioners can use 'can-do' statements right after each task or lesson. Some examples of self-assessment items driven from the lesson objectives are (1) *I could ask my partner to elaborate more on the information he/she provided*; (2) *I could rephrase what my partner said during our interaction.*

Another form of useful classroom-based assessment is peer-assessment. Along with self-assessment, peer-assessment actively invites learners to assessment practices. When designing peer-assessment, the first step is to review the course/lesson objectives and select target constructs. The constructs are the basis of criteria in a rubric or evaluation criteria (Table 3).

Table 3.   Peer-assessment examples

Questions	Likert scale
1. My partner used <target grammar> when he/she asked questions.	Absolutely not 1— 2— 3—4—5 Definitely
2. My partner asked clarification when I talked about my home country.	Absolutely not 1— 2— 3—4—5 Definitely
3. My partner rephrased what I said.	Absolutely not 1— 2— 3—4—5 Definitely

The examples above are simplified to provide an overview. VE practitioners are encouraged to specify or broaden the definitions of constructs for their assessment purposes, and the constructs should be closely related to the lesson objectives.

5.   Conclusion

Language learning is one of the major goals in VE, and it is critical to identify whether the goal was accomplished. Well-identified/developed language assessment will help VE practitioners examine the success of VE practices. In order to support VE practitioners and researchers

interested in assessing the language development of their students during VE, we first looked to prior VE research to observe some common types of assessment practices, each of which offered advantages and drawbacks. We then looked to the field of language assessment for further ideas and considerations that VE practitioners could follow in conducting language assessment. Assessment-related tasks in VE practices are daunting, and are extremely under-researched (O'Dowd, 2010). In the present paper, we addressed the dearth of information and provide practical guidelines for VE practitioners. By doing so, we attempted to enhance VE practitioners' language assessment literacy so that they can make informed decisions and take advantage of valuable information that language assessment can provide.

6.   Acknowledgments

References

Anthony, L. (2019). *AntConc* (Version 3.5.8) [Computer Software]. Waseda University. https://www.laurenceanthony.net/software

Bachman, L. F., & Palmer, A. S. (2002). *Language testing in practice: designing and developing useful language tests*. Oxford University Press.

Belz, J. A., & Vyatkina, N. (2005). Learner corpus analysis and the development of L2 pragmatic competence in networked inter-cultural language study: the case of German modal particles. *Canadian Modern Language Review/ Revue canadienne des langues vivantes, 62*(1), 17-48.

Belz, J. A., & Vyatkina, N. (2008). The pedagogical mediation of a developmental learner corpus for classroom-based language instruction. *Language Learning & Technology, 12*(3), 33-52. https://doi.org/10125/44154

Brown, H. D. (2004). *Language assessment: principles and classroom practices*. Longman.

Butler, Y. G., & Lee, J. (2006). On-task versus off-task self-assessments. *Modern Language Journal, 90*(4), 506-518.

Butler, Y. G., & Lee, J. (2010). The effects of self-assessment among young learners of English. *Language Testing, 27*(1), 5-32.

Carr, N. T. (2011). *Designing and analyzing language tests*. Oxford University Press.

Cheng, W., & Warren, M. (2005). Peer assessment of language proficiency. *Language Testing, 22*(1), 93-121. https://doi.org/10.1191/0265532205lt298oa

Cotos, E. (2017). Language for specific purposes and corpus-based pedagogy. In C. A. Chapelle & S. Sauro (Eds), *The handbook of technology and second language teaching and learning* (pp. 248-264). John Wiley and Sons, Inc.

Council of Europe. (2001). *Common European framework of reference for languages: learning, teaching, assessment*. Cambridge University Press.

Council of Europe. (2018). *Common European framework of reference for languages: learning, teaching, assessment companion volume with new descriptors*. https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989

Cummins, J., & Sayers, D. (1995). *Brave new schools: challenging cultural literacy through global learning networks*. St. Martin's Press.

EVOLVE Project Team. (2020). The impact of virtual exchange on student learning in higher education. EVOLVE Project publication. http://hdl.handle.net/11370/d69d9923-8a9c-4b37-91c6-326ebbd14f17

Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly*, *9*(2), 113-132.

Harris, M. (1997). Self-assessment of language learning in formal settings. *ELT Journal, 51*(1), 12-20.

Helm, F., & van der Velden, B. (2019). *Erasmus + Virtual Exchange 2018 impact report*. Publications Office of the European Union. https://doi.org/10.2797/668291

Inbar-Lourie, O. (2008). Constructing an assessment knowledge base: a focus on language assessment courses. *Language Testing, 25*(3), 385-402.

Lee, J. (2019). A training project to develop teachers' assessment literacy. In E. White & T. Delaney (Eds), *Handbook of research on assessment literacy and teacher-made testing in the language classroom*. (pp. 58-80). IGI Global.

Lee, J. (2020). Investigation of external and internal quality assurance strategies in online language teacher education. In H. S. Kang, D. S. Shin & T. Cimasko (Eds), *Online education for teachers of English as a global language.* Routledge.

Lee, J., & Butler, Y. G. (2020). Where are language learners? Reconceptualizing language assessment literacy. *TESOL Quarterly, 54*(4), 1098-1111. https://doi.org/10.1002/tesq.576

Lewis, T., & O'Dowd, R. (2016). Online intercultural exchange and foreign language learning: a systematic review. In R. O'Dowd & T. Lewis (Eds), *Online intercultural exchange: policy, pedagogy, practice* (pp. 21-66). Routledge.

Little, D., & Brammerts, H. (1996). (Eds). A guide to language learning in tandem via the internet. *CLCS Occasional Paper, 46*. http://files.eric.ed.gov/fulltext/ED399789.pdf

Mok, J. (2011). A case study of students' perceptions of peer assessment in Hong Kong. *ELT Journal, 65*(3), 230-239.

O'Dowd, R. (2010). Issues in the assessment of online interaction and exchange. *Telecollaboration, 2*, 337-360.

O'Dowd, R. (2016). Emerging trends and new directions in telecollaborative learning. *CALICO Journal, 33*(3), 291-310. https://doi.org/10.1558/cj.v33i3.30747

Saito, K., & Akiyama, Y. (2018). Effects of video-based interaction on the development of second language listening comprehension ability: a longitudinal study. *TESOL Quarterly, 52,* 163-176. https://doi.org/10.1002/tesq.362

Sauro, S. (2009). Computer-mediated corrective feedback and the development of L2 grammar. *Language Learning & Technology, 13*(1), 96-120. https://doi.org/10125/44170

Solano-Flores, G. (2016). *Assessing English language learners: theory and practice*. Routledge.

The EVALUATE Group. (2019). *Evaluating the impact of virtual exchange on initial teacher education: a European policy experiment*. Research-publishing.net. https://doi.org/10.14705/rpnet.2019.29.9782490057337

The Stevens Initiative. (2019). *Evaluating virtual exchange: a toolkit for practitioners.* https://www.stevensinitiative.org/wp-content/uploads/2019/09/Evaluation-Toolkit-Final.pdf

Vinagre, M., & Muñoz, B. (2011). Computer-mediated corrective feedback and language accuracy in telecollaborative exchanges. *Language Learning & Technology, 15*(1), 72-103. https://doi.org/10125/44238

Journal of Virtual Exchange (JVE) is an online, open-access, peer-reviewed journal aimed at practitioners and researchers in the field known variously as virtual exchange, telecollaboration, or online intercultural exchange. It is the official journal of UNICollaboration (https://www.UNICollaboration.org/), the international academic organisation dedicated to supporting and promoting telecollaboration and virtual exchange in higher-level education.

Disclaimer. University of Groningen Press does not take any responsibility for the content of the pages written by the authors of this article. The authors have recognised that the work described was not published before, or that it was not under consideration for publication elsewhere. While the information in this article is believed to be true and accurate on the date of its going to press, neither UniCollaboration nor University of Groningen Press can accept any legal responsibility for any errors or omissions. Additionally, the publisher makes no warranty, expressed or implied, with respect to the material contained herein. While University of Groningen Press is committed to publishing works of integrity, the words are the authors' alone.

Trademark notice. Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Copyrighted material. Every effort has been made by the editorial team to trace copyright holders and to obtain their permission for the use of copyrighted material in this article. In the event of errors or omissions, please notify the publisher of any corrections that will need to by incorporated in future editions of this article.

Typeset by Research-publishing.net (https://research-publishing.net)

Noto fonts are open source. All Noto fonts are published under the SIL Open Font License, Version 1.1. Noto is a trademark of Google Inc. (https://www.google.com/get/noto/).

Appendix A

NAME	Standards	Language	Skills	Interface	Cost	Age	Turnaround time	Website
WIDA	WIDA Standards	English	L, S, R, W	Computer-based/paper-based	$150 for 15 sets	USA PreK - 12	15 mins - 4 months	https://wida.wisc.edu/assess/choosing-assessment
TOEFL		English	L, S, R, W	Paper or Online, approx. 200 mins	$205	(Young) adults	6 days online release	https://www.ets.org/toefl/
IELTS	(CEFR)	English	L, S, R, W	Face to Face	Varies by test center	Higher Education and above	13 Days	https://www.ielts.org/en-us/
TOEIC		English	L, S, R, W	Audio, paper, written	Varies by test center	Adults	Unclear	https://www.ets.org/toeic/test-takers
TCF	CEFR	French	L, R, G, (S)	Computer-based (in-person speaking)	Varies by test center	College and older	15 days	https://www.ciep.fr/en/tcf
DELE	CEFR	Spanish	R, S	Online and in-person	Varies by test center	Adults over 16	3months	https://examenes.cervantes.es/es/dele/que-es
HSK		Mandarin Chinese	L, R, S, W	Computer and Oral	$20 and up	Primary-Professional level	Varies by test date	http://www.chinesetest.cn/gosign.do?id=1&lid=0
JLPT		Japanese	reading and listening	Online, test center	Varies		About three months	http://www.jlpt.jp/e/about/index.html
Goethe-Zertifikat	CEFR	German	L, S, R, W	Test center	€ 105- € 295	Youth-Adults	14 days after the exam	https://www.goethe.de/en/spr/kup/prf/prf.html
TestDAF		German	S, L, W	Speaking online	Varies by test center	Youth-Adults	About 6 weeks after examination	https://www.testdaf.de/
ZDfB	CEFR	German	L, S, R	Paper, online	€ 270	Adults	No longer than 6 weeks	https://www.goethe.de/de/spr/kup/prf/prf/gc1.html
CAPLE	CEFR	Portuguese	L, S, R, W	In person, paper test	€ 72 – 142	14 and above	Unknown	https://www.portugaleducation.info/tests/caple-portuguese-language-proficiency-test.html